DOCUMENT RESUME

ED 442 309 FL 801 384

AUTHOR Bonk, William J.
TITLE Testing ESL Learners' Knowledge of Collocations.
PUB DATE 2000-00-00
NOTE 69p.
PUB TYPE Reports - Research (143)
EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Applied Linguistics; *English (Second Language); Evaluation Methods; Factor Analysis; *Item Analysis; Language Patterns; Language Proficiency; *Language Research; *Language Tests; Lexicography; Second Language Instruction; Second Language Learning; Speech Communication; Test Items; Testing; Vocabulary
IDENTIFIERS *Lexical Collocation

ABSTRACT

This study reports on the development, administration, and analysis of a test of collocational knowledge for English-as-a-Second-Language (ESL) learners of a wide range of proficiency levels. Through native speaker item validation and pilot testing, three subtests were developed and administered to 98 ESL learners of low-intermediate to advanced proficiency. Descriptive statistics and reliability estimates for the test administration are calculated, and the characteristics of the test items, subtests, and response modes were examined using traditional item analysis, Item Response Theory, and generalizability theory methods. Two of the three subtests were found to perform well as norm-referenced measures of construct, and areas for further testing and research were pinpointed. Observed collocational knowledge was found to correlate strongly with a measure of general ESL proficiency, while length of residence had negligible predictive power of collocations on test performance. Exploratory factor analysis revealed that collocations items tended to load on different factors from general proficiency items, giving preliminary evidence of construct validity. Extensive references and quantitative data, displayed in various scatter diagrams, bar charts, and other graphics and figures, are also provided. (Adjunct ERIC Clearinghouse for ESL Literacy Education) (Author/KFT)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 309

Running Head: Testing Collocations

Testing ESL Learners' Knowledge of Collocations

William J. Bonk

U.S. State Dept. EFL Fellow

BEST COPY AVAILABLE

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

BONK

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1 2

L801384

ABSTRACT

This study reports on the development, administration, and analysis of a test of collocational knowledge for ESL learners of a wide range of proficiency levels. Through native speaker item validation and pilot testing, 3 subtests were developed and administered to 98 ESL learners of low-intermediate to advanced proficiency. Descriptive statistics and reliability estimates for the test administration are calculated, and the characteristics of the test items, subtests, and response modes were examined using traditional item analysis, Item Response theory, and generalizability theory methods. Two of the 3 subtests were found to perform well as norm-referenced measures of the construct, and areas for further testing and research were pinpointed. Observed collocational knowledge was found to correlate strongly ($r = .73$) with a measure of general ESL proficiency, while length of residence alone had negligible predictive power of collocations test performance. Exploratory factor analysis revealed that the collocations items tended to load on a different factor from general proficiency items, giving preliminary evidence of construct validity.

Testing ESL Learners' Knowledge of Collocations

Introduction

Native speakers have extensive knowledge of how words combine in their language, and they use this knowledge when they retrieve lexical items and link them appropriately in language production. Systematic use of these combinations is considered an important element of native speaker competence (e. g., Pawley & Syder, 1983; Ellis, 1996, and, in the case of second language (L2) learners, of native-like L2 production [McCarthy, 1990]). Such recurrent combinations of lexical items are often referred to as collocations or formulaic speech in the linguistics literature, though there is widespread variation in the usage of these terms. While some research has looked at the role of unanalyzed chunks and formulaic speech in second language acquisition (e.g., Peters, 1983), the use and development of this domain of language knowledge among adult second language learners has remained anecdotal in nature and for the most part unresearched. The development of reliable and valid measures of this construct are perhaps a first step towards a more complete understanding of its importance in L2 use and acquisition.

Lexical Knowledge

Native speakers (NSs) possess richly detailed knowledge about lexical items in their language, such as various types of "meaning," abstract semantic information, connotations, and receptive and productive knowledge of conventional expressions containing particular words, to name only a few. While in the past a great deal of linguistic speculation and research (Irujo, 1986) focused on speakers' knowledge of the relatively colorful expressions and idioms (e.g., kick the bucket), more mundane lexical combinations have only recently become an object of attention. This pattern of research interest has perhaps been detrimental to a general understanding of the scope

of the topic and its importance in language production, since there is some evidence (Howarth, 1996) that idioms and frozen form expressions are relatively infrequent (approximately 5% of total text) in native speaker academic writing, while restricted collocations as defined in this study are much more prevalent (34%) [p. 122]. One further barrier to study has perhaps been the morass of overlapping terminology used by various researchers over the years to describe this and related areas of lexical investigation, such as "prefabricated routines," "gambits," "colligations," "lexicalized sentence stems," "formulaic speech," "prefabricated patterns," and "polywords." Overall, there seem to be three common usages in the literature for the term "collocation," which will be considered in turn.

Definitions of the Term "Collocations"

Much recent work on collocations has emerged from or been influenced by corpus-based research (see Benson, Benson, & Ilson, 1986; Kennedy, 1990; Aijmer & Altenberg, 1991; Sinclair, 1990; Kjellmer, 1995; also Oppenheim, 1993, for somewhat similar treatment of formulaic speech). In general, these researchers purposely adopt a broad interpretation of the term collocation, giving this designation to any recurrent pairs or groups of words which emerge from the corpus with a greater frequency than could be predicted by their individual frequencies as lexical items. This definition is therefore not a strictly linguistic one, but is rather a practical, operational one, reflecting the procedure used to extract these items from the corpus.

A second commonly encountered use of the term "collocations" in recent literature (e.g., Ellis, 1996) is a general linguistic one which seems to denote any polyword structures or recurrent sequences of language. This is similar to the definition used by corpus linguists as described above, but it is not restricted to the recurrent sequences in a given corpus, since it is used to talk about the phenomenon in

general rather than a way of extracting them from language data. This understanding of polyword phenomena is perhaps most often associated with research such as that of Nattinger and DeCarrico (1992), who use the term "lexical phrases" as their general designation for multiword linguistic phenomena, and suggest that conventionalized, prefabricated chunks of language are extremely common in fluent speech and writing, and that they are an important source of linguistic material for language learners to later analyze and derive syntactic and lexical information from.

Still other researchers reserve the term collocation for a much more specialized linguistic phenomenon. Howarth (1996) limited "restricted collocations" to the following: institutionalized combinations of lexical items which lie somewhere between frozen form and semantically opaque pure idiomatic phrases and free combinations of lexical items, in which one element is used in a non-literal sense, and which do not permit many substitutions on the continuum of productivity. The phrase to catch a cold would be a restricted collocation by this definition, since (a) it is immediately recognizable as a conventional phrase; (b) it uses one element in a specialized way (catch here is a somewhat figurative usage of the verb which differs from its prototypical meaning); (c) this element has a limited range of collocates (in this case, illnesses); and (d) the phrase is semantically transparent. The phrases to catch a butterfly and I didn't catch that would not be restricted collocations by this definition, since they are free combination and an idiomatic usage respectively.

The terms collocation and formulaic speech are often used interchangeably in the literature, a fact which is perhaps more due to the divergent definitions of collocation than a similarity of the various linguistic behaviors. From a theory point of view, I would tend towards a linguistically based definition of collocation such as Howarth's (1996), but argue that the term collocation is best understood as

connections between items in the mental lexicon based on lexical and semantic characteristics, and not as a chunked storage and production strategy per se, as formulaic speech may prove to be, nor as a kind of structural rule. In other words, from the fact that there are combinations of words which occur frequently in the language, and that some seem to be stored as lexical units (Aitchison, 1987), it does not necessarily follow that all word combinations are stored in this way, or that they all have some similar underlying psycholinguistic reality. Even a division of collocation into lexical and grammatical types as appears in Benson et al. (1986) may not be an entirely valid one. While it is beyond doubt that some of the 26 "grammatical collocation" types in their BBI Dictionary exist in English as complementation structure rules,[1] there is some question as to how much these have in common with the lexical collocations also included. Again, the fact that computers are able to extract significant recurrent sequences of lexical items in a corpus does not necessarily mean that all these sequences are a product of the same underlying psycholinguistic storage or language production mechanisms. Systematic research into the semantics and psycholinguistics of collocation and other types of phraseology (Howarth, 1998) seems to be lacking in most discussion of the topic (including here), and this may be a fruitful area for future investigations. However, for the purposes of this paper, the somewhat non-technical but commonly known label collocations is used for convenience's sake, while it is recognized that this label may be somewhat misleading.

Importance of Collocations and other Multi-word Linguistic Phenomena

Language users' knowledge of collocational relationships and of habitual combinations of lexical items in general has not been systematically researched in applied linguistics, despite the fact that it probably has great importance for many

aspects of language competence, most importantly in speech production. It is clear that some sort of knowledge base of how words combine is frequently accessed during language production, since certain lexical items select for others to appear (e.g., a belief in life after death, where the word belief requires in as its preposition). This type of knowledge is consequently essential for grammatical accuracy (in the broadest sense of the term). Knowledge of collocations must be of importance for the construction of utterances, since developed and routinized collocational knowledge probably means less reliance on "creative construction" in grammar and lexis, and accordingly less attention and processing, and greater fluency; this does not appear to have been the focus of any L1 or L2 research thus far.

Idiomaticity in a speech community is also dependent upon targetlike lexical knowledge. Nativelike selection (Pawley & Syder, 1983) means among other things that speakers or writers are able to choose and recognize appropriate vocabulary and expressions for the social situation and register (Howarth, 1996). Conventionalized language in appropriate amount and accuracy gives speakers the impression of control and fluency, while a lack or overuse of it can make a text seem very "accented" (Yorio, 1989). The acquisition of appropriate collocations (e.g., administer a test) would appear to be an essential part of acquiring and demonstrating a competence in that speech community, since it reflects a deep knowledge of the common lexis of the field.

Language comprehension is also a likely area where the effect of collocational knowledge has potential importance. All current models of speech processing recognize relatively powerful "lexical effects" whereby lexical recognition is influenced by linguistic environment, although they make various claims as to the point at which higher-level information becomes accessible to a listener (Frauenfelder

& Tyler, 1987). Interactionist models of listening comprehension, for example, describe how listening can involve sampling the sound signal and matching it with expectations, rather than the careful hearing and identification of each morpheme (Rost, 1994). Quick, top-down-aided processing of language would probably be problematic without knowledge of habitual and frequent patterns in that particular language in the form of conventional word pairings and multi-word phrases. Access to this type of knowledge may significantly reduce the amount of work a listener or reader has to do, since lexical access can occur without focused attention on all aspects of the stream of speech. The use of frequently occurring word combinations may also help an audience to more immediately understand an attempted message when they experience difficulties in decoding it due to the presence of non-target-like sound shapes, such as in the speech of a NNS. Conversely, unconventional expressions or collocations may just as well cause a listener or reader to hit "bumps" and experience problems in the comprehension of the text. This is also an area in which no research seems to have yet been attempted.

Problematicity of Collocations for NNSs

Beginning and intermediate learners may not have much available processing capacity to pay careful attention to how words are conventionally combined in speech or in a written text. As Howarth (1998, p. 162) points out, it may also be unclear to them how restricted a given collocation is. This may result in a complete avoidance of non-free combinations of words, or conversely in a significant foreign "accent" in their L2 production, due to the presence of many unconventional collocations; for most learners it is probably a combination of both these strategies. As in the case of phonology, a strong foreign "collocational accent" could give interlocutors a misguided impression of one's competence in the L2, and influence the type of input

one receives from native speakers. It has also been pointed out by various researchers (Howarth, 1996; Brown, 1974) and suggested in a small-scale study (Zimmerman, 1993) that language instructors themselves are not often aware of the concept of collocation, and consequently may not be drawing students' attention to it in their instruction, even if it is present in classroom teaching materials.

Unfortunately, the bulk of the research on NNS knowledge of word combinations has centered on true idioms rather than the more productive areas of the restricted section of the idiomaticity cline (e.g., Irujo, 1986), and NNSs' proficiency (or lack thereof) to form acceptable collocations is only now beginning to be systematically researched. It may be that even among the best language learners, those completely native-like in their grammar and pragmatics, low-frequency lexical items and restricted collocations will always present problems; indeed, lexical phenomena (and of course phonology) may be the only remaining readily perceived non-native-like aspects of their language production. This is of course to be expected, given the number of potential errors and the haphazard way in which this knowledge must be acquired. Because there are few generalizations that one can make about the collocational restrictions in the language (there are no general rules to follow), learning or teaching them in a systematic, time-saving way seems an impossible task. As Howarth (1996) points out, "Learners are, understandably, generally unaware of the large number of clusters of partially overlapping collocations, which display complex semantic and collocational relationships. It is, of course, not only learners who are unaware of this category: it is an area unrecognized in language pedagogy and little understood in lexicography" (p. 162).

It therefore appears that the task of acquiring native-like collocational knowledge in an L2 is a long and difficult one. Researchers and teachers working in

this area have long spoken of learners' inadequate proficiency to produce acceptable collocations in a foreign language (Brown, 1974; Richards, 1976; Pawley & Syder, 1983; Riopel, 1984; Mackin, 1986; Bahns, 1993; Zhang, 1993). At this writing, however, few attempts have been made to investigate L2 learners' actual collocational proficiency in any language, and there is a particular lack of studies involving a wide variety of proficiency levels.

Collocations and L1 transfer

A number of researchers have tested second language learners' knowledge of lexical collocations with an emphasis on the role of the L1 in creating transfer of forms from L1 to L2. Hussein (1991), Marton (1977), Bahns and Eldaw (1993), and Biskup (1992) have reported studies testing homogeneous L1 groups of EFL students on cloze and L1-L2 translation-type items. All these studies have used verb-object restricted collocations as the basis for their tests. These researchers have consistently found that learners commit many errors in such tasks, and that they are highly likely to transfer restricted collocations from the L1 to the L2 when they are not sure of the correct L2 form. The researchers recommend contrastive analysis and corresponding pedagogical intervention in order to further students' knowledge of the target language forms. At this writing it does not seem that a study of this type has yet been attempted, despite that fact that it might give interesting and potentially useful results. Unfortunately, the studies cited above do not provide necessary information regarding the general proficiency level of the examinees, or statistical information on the test instruments themselves, so it is somewhat difficult to know exactly how solid their findings are. Nevertheless, it seems entirely plausible that L1 transfer could play a large part in the production of second language collocations when there is a knowledge deficit, and that this might be a reflection of a general hypothesis of

lexical similarity as a production strategy, as long as the figurative sense of the collocate does not seem to be too far from its core meaning (cf. Kellerman, 1986).

General Collocations Testing Studies

There have been few published studies measuring the collocational proficiency of ESL learners, and none in L2s. In order to investigate the correlation between general English proficiency and collocations knowledge, Ha (1988) measured ESL learners' collocational knowledge on selected response cloze-type tests. Three types of collocations (verb-preposition, verb-object, and adjective-noun) were selected, and items were developed by consulting the BBI Combinatory Dictionary (Benson et al., 1986); a cloze test was also administered to measure general proficiency. Ha attempted to include both low- and high-frequency collocations in test items (in order to control for frequency in the input) by soliciting NS metalinguistic judgements as to the relative frequency of the collocations. The test instruments used had reasonably good reliability estimates (cloze K-R 21 = .86 .82, .73 and .70 respectively for each of the collocations subtests), and a robust correlation (r = .83) was found between collocation measures scores and general proficiency. The two measures which were correlated may have been confounded in the study, however, given the similarities between item types in the collocations and the cloze (proficiency) tests.

Gitsaki (1996) conducted what is perhaps the largest study of learners' knowledge of collocations. Gitsaki tested 275 adolescent Greek schoolchildren's ability to produce English collocations, investigating the accuracy and frequency of students' free production of 37 types of collocation (the 26 grammatical and seven lexical collocational patterns from Benson et al. [1986], plus four additional types of lexical collocation suggested by Zhang [1993]) in essays, as well as their performance

BEST COPY AVAILABLE

on blank-filling and L1-L2 translation tests. She found that the accuracy and frequency of their use of types of collocations increased with their proficiency (as defined by six types of analyses of the language found in the essays) on both the blank-filling and translation tests, and that there was some evidence of a pattern of development of knowledge of collocational types in the form of an implicational scale. Students' free production of collocations in the essays, however, yielded mixed results: between-group differences were generally not in the expected direction (for example, Gitsaki reported that learners in the middle proficiency group produced more frequent and accurate adjective-noun collocations than learners in the higher proficiency group), suggesting that the interaction between level of proficiency and use of collocations in the second language is somewhat more complex than objective test results might indicate.

Some methodological issues in this study, however, may have a bearing on the interpretation of Gitsaki's findings. Instead of determining learners' proficiency levels and grouping them using an independent measure, as would be the normal procedure in a testing study, three intact groups were used (students in three successive years) and the groups' essays were analyzed for six measures of proficiency (holistic rating, TLU of articles, lexical density, words per T-unit, error-free T-units, and S-nodes per T-unit). Statistical tests were used to determine if differences were significant between the groups on these six measures; there were significant differences between groups on five of the six measures, but not always in the expected direction. All in all, though there seemed to have been some differences between the groups in proficiency, it is not clear how great these differences actually were; generally speaking, it is unlikely that this population represented a wide range of abilities. Additionally, the essays which were analyzed to determine general English

BEST COPY AVAILABLE

proficiency level were the same ones later measured for free production of collocations; this confounds the two variables in the study, since there may have been some interaction between use of collocations and the six proficiency measures described above. Furthermore, learners in each of the three proficiency levels were not given the same items on the blank-filling and translation tests; Gitsaki (personal communication) intended to measure collocation types rather than collocation items, but in doing so did not take into account item difficulty within the same collocation type. Therefore we cannot be sure that a higher score on the test items necessarily reflected higher levels of collocational knowledge. Finally, reliability estimates and item analysis results were not reported for the collocations or proficiency tests, making it unclear if the testing instruments were functioning well as measures of either of these constructs.

Assuming that these methodological problems do not invalidate the results, Gitsaki's study found a positive relationship between general proficiency and collocational knowledge, and perhaps even some sort of developmental pattern, whereby learners at higher levels of proficiency tend to use certain types of collocations more often and more accurately than others - namely noun-preposition and adjective-preposition collocations. It does not give us good information about how well tests of the construct perform, however.

Because collocations testing had thus far been conducted in a somewhat unsystematic fashion in the literature, without consistent adherence to common test development practices and without detailed item analysis or consideration of test reliability and validity, a new study using a carefully developed and analyzed test seemed justified in order to address testing concerns and to determine the relationship

BEST COPY AVAILABLE

between collocations knowledge and more controlled measures of language proficiency. The following research questions were posed:

1. How reliable is the collocations test and its subtests for the targeted population?
2. Do the item development procedures used result in items of good discrimination?
3. Is there a correlation between proficiency in producing and recognizing collocations and general English proficiency?
4. Is there a correlation between proficiency in producing and recognizing collocations and length of residence (LOR) in an English-speaking environment?
5. Do lower-proficiency learners demonstrate any knowledge of collocational relationships?
6. Can evidence of validity for the collocations test be shown?

Method

Development of the Collocations Test

A pilot test to measure NNS proficiency in English collocations was developed using methods described as follows.[2] Items of the three types were targeted for inclusion in the test: verb-object collocations, verb-preposition combinations, and figurative-use-of-verb phrases. Sixty preliminary items were written (20 in each of three subtests) with special care taken to separate collocating elements syntactically (e.g., I took lots of pictures rather than I took a picture), and to use verbs in various configurations such as in present and past tenses, gerunds and plain forms, in affirmative and negative sentences, and in active and passive modes, in order to tap into learners' more complete knowledge of these forms, rather than merely their memorized knowledge of unanalyzed chunks. The rationale for including these three collocation types was that they had been used in earlier experiments (e.g., Bahns,

1993; Ha, 1988), had been labeled as collocations types by lexicographic analysis (Benson et al., 1986), but their status as types of similar knowledge of word combinations (as opposed to phrase structure rules) is indisputable.

Native speaker volunteers provided baseline data by taking this 60-item pilot version of the collocations test. Informants were chosen for participation based on the following criteria: non-language-teaching professionals, from the mainland of the U.S., five male and five female. When tests had been completed by all informants, results were compared and only those items upon which there was unanimous agreement among the 10 NSs as to the correct answer were retained.[1] This process resulted in a 30-item pilot test: three subtests of 10 items each. Later inspection of the distribution of pilot test scores indicated that examinees at the lowest levels of collocational proficiency may not have had many items within their reach, so 20 new items were developed and added to the original 30-item test prior to the main test administration. These underwent the same item validation procedures as described above, again with 10 non-language-teaching NS informants.

Other Materials

To measure general proficiency in written English among NNSs, a version of the TOEFL (based on an actual past version of the test) was condensed by eliminating the listening section and reducing each of the other sections; this made the proficiency test 49 items long, with an appropriate amount of time allotted for each section. A biodata questionnaire was also developed to gather information regarding subjects' age, gender, nationality, native language(s), age of first daily contact with English,

length of residence in English-speaking countries, and amount of formal instruction in English.

Participants

Sixty-two NNS volunteers (21 males, 41 females) participated in pilot testing of the collocations and proficiency measures. In subsequent main test administration, 98 adult NNS's (41 males and 57 females) from the same population took the test. The majority (87%) were of East-Asian first languages, and their English proficiency varied from low intermediate to very proficient advanced users of English (as indicated by the distribution of scores on the general proficiency tests in this study; see Figure 5). All examinees were students at the University of Hawai'i, and therefore had adequate English reading and writing skills and familiarity with the TOEFL to be able to take the tests, since TOEFL scores were required for application to the university. Subjects were mixed instructed-naturalistic learners of English (all had had many years of ESL instruction as well as some experience interacting in the target language) whose first experience living in an English-speaking country was after the age of 13.

Procedures and Scoring

Both collocations and proficiency tests were administered in the same order during class time to existing groups of students. While some examinees left answers blank, unanswered items were generally not concentrated at the end of each subtest, suggesting that subjects had had enough time to read all questions and answer those that they felt capable of attempting. A complete administration of both tests took only 60 minutes, so fatigue was probably not a major factor in the subjects' scores. In order to check for test fatigue, however, after the test administration item facility

BEST COPY AVAILABLE

values were correlated with item numbers using the Pearson product-moment correlation, and the resulting value was not significant at $p < .01$.

Tests were scored by hand by the researcher. Examinee names and other biodata were not evident at scoring time, and answers for the blank-filling data were counted as correct if they matched native speaker pilot test responses; spelling and grammar errors were not counted incorrect responses, as long as a recognizable facsimile of the correct lemma was supplied.

Analysis

Pearson's product-moment correlation was used to compare student scores on each test and subtest, and collocations scores and length of residence (LOR) at an error level of $p < .05$, one-tailed test. Descriptive statistics were calculated for all subtests and total test scores. Traditional item analysis was done in order to observe the performance characteristics of the various items and subtests. Test reliability was calculated using the K-R 20 formula. Simple regression analysis was applied to the means of the collocations and proficiency test scores to observe the ideal line of regression. Collocations test data were analyzed using a one-parameter IRT (Item Response theory) model to estimate item parameters and evaluate examinee and test performance using these parameters. Collocation test score data were also analyzed using generalizability theory in the form of a two-facet $p \times (i:s)$ design in order to identify sources of error in the test, to estimate the generalizability of the test scores in this administration, and to estimate the efficiency of other potential configurations of subtests and items. Exploratory factor analysis was performed on all test data in order to investigate the convergent and divergent validity of the instrument used in measuring collocations.

Results

BEST COPY AVAILABLE

Test Descriptive Statistics

Descriptive statistics for the collocations test and its subtests (see Table 1) and for the proficiency test and its subtests (see Table 2) were calculated . Only the verb-preposition collocations subtest had a non-normal distribution (see Figure 3) and an unacceptably low K-R 20 reliability coefficient (.47) for the same number of items as the other subtests. The other subtests (Figures 2 and 4; Table 1) were normally distributed, well-centered, and had reasonably high reliability coefficients considering the population and test size[1]. Overall collocations test reliability was estimated at .83; given the fact that this was an unimproved version of the test, the revision of items, prompts, and distractors would likely yield a test of very good reliability.

The proficiency test data were less normally distributed than those of the collocations test (see Figure 5) and displayed some measure of negative skewness, which is to be expected given the presence of many advanced NNSs of English in this subject pool who were able to "max out" the test. Nevertheless, it was generally a reliable measure of proficiency (K-R 20 = .85) for this population. Descriptive statistics for the proficiency subtests are presented in Table 2.

Item Analysis

Item facility, item discrimination, and point biserial coefficients were calculated for the collocations items (see Table 3). It is clear that, through the item development and validation procedures detailed above, it was not difficult to generate a large number of apparently good, well-discriminating items as shown in Table 3 above. Even some multiple-choice items (fig. - verb types) whose distractors had never been revised showed promise as good items.

Collocations Correlational Data

Pearson product-moment correlations were performed on collocations total test scores and students' self-reported LOR for the pilot administration only; this relationship was .39, a statistically significant value at $p < .05$, one-tailed test. The LOR data were not normally distributed, and examination of the scatterplot established that the relationship was not linear in nature, suggesting that the data did not fit well into this sort of statistical model. This analysis was therefore not undertaken for the 98 participants in the subsequent test administration.

Pearson product-moment inter-test correlations were calculated for all subtests and for total scores; these results are presented in Table 4. The correlation between collocations test and proficiency test mean scores was .61, indicating a shared variance (coefficient of determination) of .37.[7] After correction for attenuation, necessary because of the unreliable amount of variance in each measure (Hatch & Lazaraton, 1991), the correlation is $r_{(A} = .73$, $r^2 = .53$. While all values in the correlation matrix were significant at $p < .05$, one-tailed test, they are all in a similar range with none particularly standing out as a high or low value.

Regression Analysis

A simple regression analysis model was fitted to the collocations and proficiency test score data after data were checked for violations of the assumptions of this statistic based on Neter, Wasserman, and Kutner (1990). The scatterplot of the simple regression (see Figure 6) shows this relationship and the ideal line of regression between collocations and proficiency mean scores. The relationship is basically linear in nature, although there is obviously a great deal of error in this correlation.

Rasch Analysis

BEST COPY AVAILABLE

A two-parameter IRT model (Rasch analysis) was fitted to the results of the collocations test. This analysis was performed using the BILOG software package (Mislevy & Bock, 1992), with all items were entered as a single test. Item thresholds and reliability estimates are presented in Table 5, along with the error estimates associated with the threshold values; item fit statistics are in Table 6. An item-to-person fit map is presented in Figure 7, and a chart of information statistics in Figure 8.

As did the traditional item analysis, Rasch item analysis (Table 5) indicates that as a whole the test of collocations seems to have performed reasonably well with this subject population. There is a good mix of item threshold values, and the errors estimates associated these threshold values are low relative to those on items of very high or low difficulty, whose parameters tend to be more difficult to estimate because of more limited data at the ends of the ability scale. Slope values (Table 5) were by and large quite high, showing the effective discriminatory power of this collection of items.

BILOG provides reliability estimates for each item (calculated by xxx, see Table 5) generally fall into a range of .15 to .30. Average reliability was .23, .16, and .20 by subtest respectively, providing more evidence that the second subtest was problematic compared to the other two. These are directly related to the amount of maximum information, also shown in Table 5.

Four items (3, 27, 28, and 46, see Table 6) had chi-square probability values of less than .05 indicating that they may not fit the model well, although this dataset is far too small to provide reliable estimates of item or candidate misfits (Hambleton et al., 1991). The BILOG manual (Mislevy & Bock, 1990) suggests that values of.01 and under indicate significant deviation from the model and a need for revision.

BEST COPY AVAILABLE

It should be noted that the threshold value corresponds to the point on the logit scale of maximum item-level information (Table 5), and that items yielding good amounts of information should have threshold values distributed throughout the logit scale in order to make the test work well in discriminating examinees at a wide range of levels. The map showing fit of items to individuals (see Figure 7) provides evidence that the majority of examinees, even lower levels (those below -1.0 on the logit scale) have a quantity of items which match their ability level, and that therefore the test should be able to discriminate among them, if it is at all possible to do so by testing them on this type of knowledge. The information and error map below (Figure 8) confirms that good information is available within 2 standard deviations on either side of the mean for this population. Overall IRT-based reliability for this administration was estimated at .93.

Generalizability Analysis

Because of the exploratory nature of collocations testing, Generalizability theory (G-theory) analysis was also applied to collocations test scores in the form of a two-facet design (person, items nested within subtests, or p x (i:s)) in order to further investigate the nature of the three subtests and two response modes used. One randomly selected item was dropped from each of the first two subtests in order to balance the model at 16 items per subtest. Table 7 provides estimates of the variance components associated with the facets included in the model for this test.

The variance component associated with between-person variation (.0183) is high relative to the others, indicating that most variance in the test is explained by differences in ability on the construct rather than by characteristics of the test method itself. The low subtest (0) and person-by-subtest components (.0004) suggest that varying the number of subtests will not tend to increase the reliability of this

instrument, if they are similar to the types of item sampled in these subtests. The amount of variance contributed by the items themselves (represented by i:s) is low relative to person variance p, as is the overall interaction between persons and items persons (p x (i:s)). The sum total of test error (variance components other than person variance, or Δ) is .0052, or 22% of the total .0235. This can be interpreted to mean that, though there is some non-systematic error in the test, it accounts for no more than 22% of the total variance, and that the test is generally internally consistent. Test improvement measures would likely bring this amount of error to even lower levels.

Next, generalizability coefficients (G-coefficients) were calculated for varying configurations of a hypothetical collocations test using similar subjects and item types in a D-study. The results of this analysis are presented in Table 8, organized by effect on the G-coefficient. Upon examination of Table 8, it is clear that there is a more or less arithmetic relationship between total number of items and G-coefficient for this particular test model. As already noted, subtests do not have much effect in this model; comparing the actual administration using three subtests of 16 items each to potential ones using two subtests of 24 items each or six subtests of eight items each, changes in the G-coefficient are negligible. In this particular case, since the verb-preposition subtest has already been under suspicion of not adding much discrimination to the test, eliminating it and leaving 16 items in the other sections would theoretically yield a G-coefficient of .76, which might be acceptable. Eliminating non-discriminating items from the remaining subtests and replacing them with better ones would likely increase this further, and make the overall test shorter and more reliable compared to the full test.

Factor Analysis

Factor analysis was performed on the three collocations subtests and the three proficiency subtests with the Eigenvalue set at 1.0. Communality values were inspected, and were sufficiently high (see Table 9) to conclude that the variables were well-defined by the solution, and that there were no outlying variables. Two factors were extracted in the solution (see Table 9). Because a significant correlation was expected to exist between the variables, an oblique solution was the appropriate one. The factor loadings in the oblique solution in Table 9 display a clear pattern of convergence of collocations variables on Factor 1 and proficiency variables on Factor 2. Factor 1 seems to be the factor which reflects knowledge of lexical relations, while Factor 2 appears to be a facet more related to general language proficiency. The direct variance contributions (see Table 10), representing the amount of total variance each factor accounts for individually, are high, indicating that each factor contributes a great deal of unique variance to the solution (41% and 39% respectively). It is interesting to note that there is an overlapping (or joint) contribution on Factor 1. This may be interpreted to mean that there is some contribution of proficiency to the collocations factor, but that it is small (20%) compared to the influence of collocations knowledge on this factor. The results of this factor analytic solution suggest that the collocations and proficiency subtests are measuring quite different things, and constitutes some preliminary evidence for the construct validity of this test.

Discussion

Test Reliability and Item Analysis

Given that this was an unimproved version of the test (in that no poorly performing items or distractors were dropped after pilot testing, items were simply added), it appears that items with acceptably high item discrimination values can be

T COPY A

rather easily developed for the first and third collocation types. Note that while IF values are similar (Table 3), the mean of the ID and pbi correlation coefficients for the second item type (verb-prep collocations) is noticeably lower: this is the same subtest which displayed notably lower reliability (K-R 20 = .47) than the others and which was not normally distributed. While this comparison is not necessarily statistically valid, if we suppose that the items included in this subtest are a somewhat reasonable sample of the domain, then this suggests that at least among NNSs, this aspect of English proficiency is much less easily tested in this fashion. In fact, if this entire subtest is eliminated and the K-R 20 reliability for the remaining two collocations subtests is recalculated, a reliability coefficient of .79 obtains, indicating that this entire section of 17 items may contribute virtually nothing to the internal consistency of the whole test. If all items with ID's of less than .30 are eliminated from this administration, a total of 30 items remain. Recalculation of the reliability estimate of this hypothetical administration gives a reasonably high (and nearly identical) K-R 20 estimate of .82, in spite of the 40% decrease in total number of items.

Overall, this measure of English collocations demonstrated a moderately high level of reliability (K-R 20 = .83) for this group of subjects. Since this was an experimental, unimproved version of the test, it is likely that a simple test improvement measure such as the replacement of items of low discrimination with better ones would increase this figure to more acceptable levels. However, it does seem to be the case that, at least for this population of learners, some types of collocations may be less reliably tested than others. While the cloze-type production response mode proved relatively reliable (K-R 20 = .69) for verb-object collocations, this same response mode was much less reliable for the verb-preposition collocations

BEST COPY AVAILABLE

(K-R 20 = .47), suggesting that it may have been the content of the items themselves rather than a method effect which was responsible for this difference in reliability (but see a fuller consideration of this issue below). Nevertheless, the verb-object subtest performs seems to perform reasonably reliably for its size.

A qualitative examination of a sample of 24 test answer sheets (25% of the total) was undertaken in order to see if any information was there on student responses to the fill-in-the-blank type items in the verb-object and verb-prep subtests. The response data showed that examinees seemed to understand the prompts and enter semantically appropriate responses in the blanks the great majority of the time for the verb-object subtest. Out of these 408 possible responses (17 items from each of 24 test forms), 17 (4%) were left blank and 20 (5%) were of at least approximate semantic appropriateness[5]. The rest of the responses seemed to demonstrate understanding of the prompt itself and of the cultural schema being activated (for example, that chocolate is said to "spoil one's appetite;" incorrect responses included "break" and "destroy"). Test improvement in the case of this subtest would likely involve the development and pilot testing of more productive items as well as experimentation with less culturally bound concepts.

The verb-prep subtest demonstrated unacceptably low internal consistency estimates (K-R 20 = .47). An analysis of 24 randomly selected student response forms (see Table 11) was undertaken to see if any pattern could be identified in their responses. Perhaps, although ostensibly a production item, phrasal verb collocations cloze-type items as presented here actually function more like selected response items, since a limited number of possible answers is involved; this may in turn contribute to this subtest's lack of reliability in that guessing is involved. Furthermore, these students seem to know that <u>out</u>, <u>up</u>, <u>on</u> and <u>off</u> are very common

particles, and chose them far more than would be expected if they were choosing from the full spectrum of prepositions. Since the categories are so limited, it seems possible that guessing from among these high-frequency particles on unknown items was a strategy adopted by some of these examinees.

It is not clear from this data whether L2 learners acquire phrasal verbs of this type as simply memorized units, or if they perceive any of the semantic or aspect content of the prepositions in them. It is perhaps the often elusive shades of meaning that serve to confuse NNSs; indeed, even NSs are probably unable to explain exactly what prepositions in such lexical relationships mean. It seems that there is a combination of semantic, syntactic, and lexical knowledge in these expressions that makes them hard to acquire. One can only speculate as to what the above error patterns mean; however, it appears that most of the error responses in Table 11 are errors reflecting some sort of target language knowledge. The most common errors, those produced by many candidates independently on item numbers 18, 22, 26, 27, 28, and 29 for example, may indicate learner awareness of common target language phrasal verbs, and or some semantic knowledge of the preposition they chose to use. The most common error responses on item #28 formed the very common phrasal verbs to look up and to look through, both commonly associated with written texts, as suggested by the prompt. Also, although they often were not able to produce the same forms as NSs, examinees were generally not using completely inappropriate prepositions; this explains why we do not find under, apart, or back in the incorrect response list for item #7. On the other hand, items 19 and 34 had very high IF figures and no incorrect responses from these examinees. They may be commonly included in a list of phrasal verbs to be studied in school, and may be more common in speech as well, which would explain their relative ease as items. This pattern may also

provide a clue as to why this subtest is more unreliable, namely that differential instruction in various institutions may cause vocabulary items such as these to be non-scalable, as opposed to items encountered and acquired more or less haphazardly. In any case, unless they are to be tested in some more effective way, the prepositions in phrasal verbs do not seem to be an extremely useful type of item to include in a test of collocations. They are methodologically complex, and do not seem to work well as reliable and discriminating test items. They may work more as units than as combinations, and could therefore be a different linguistic phenomenon with a different psycholinguistic reality.

In terms of the reliability of particular item types, true selected response items (used in the figurative verbs subtest) performed somewhat poorly (K-R 20 = .61, see Table 1) in terms of their reliability, although given the subtest length, this is not entirely unacceptable; it is not known what contribution response mode, collocation type, or individual item characteristics make in producing such a reliability figure, since there is no corresponding subtest using similar response mode and different items types to compare it to. Distractor analysis revealed only seven answer choices not attracting any candidates' responses, so the great majority of distractors seemed to be functioning adequately. As for the other subtests, standard test improvement procedures such as the replacement of non-performing items and the introduction of better distractors in some cases would likely have a positive effect on the reliability of this subtest. Again, since this administration involved an unimproved version of the test, somewhat low reliability values are not necessarily indicative of a basic flaw in this type of item, but rather a starting point for test improvement measures.

It would of course be interesting to know the relative difficulty of each collocation type included in this test. Mean IF values were .51, .48, and .53

BEST COPY AVAILABLE

respectively, but these are not valid measures of item type difficulty unless they are randomly sampled from a large number of items for comparison. A better way of making this comparison is to use the Rasch analysis results, since the item threshold values were all put on the same scale by the computer program. A comparison of mean item threshold values for each subtest (expressed in logits in Table 5) may give us a better idea of the absolute difficulty of these types. The mean values (on the logit scale) for these three subtests are as follows: .052, -.171, and .128, respectively, indicating that the verb-prep subtest was the easiest; the verb-object subtest was in the middle, and the figurative use of verbs subtest was the most difficult. The main problem with this analysis is that all three subtests under scrutiny used different response modes: there is a substantial possibility for guessing in the third subtest and some possibility in the second, a fact which is not controlled for in the model. If we assume that guessing accounted for a significant number of correct responses in this third subtest, and we already have seen that most responses on the first and second subtests were semantically appropriate, then it is clear that the figurative verb subtest was potentially much more difficult than the other two. This interpretation seems intuitively correct upon examination of the items in this subtest, which seem to be highly idiomatic and of lower frequency than ones in the other subtests. The ultimate determination of collocations type difficulty must be decided in a study designed to test this directly, however. It may be that this subtest type would be easier and more reliable if examinees had to choose the correct sentence, as opposed to choosing the incorrect one from among three correct ones, since finding the correct response in the former might require more knowledge than in the latter.

Collocations - Proficiency Correlation and Regression

BEST COPY AVAILABLE

In this study, a moderately high level of correlation (r_{xy} = .73) was found to exist between the proficiency measures and collocational proficiency. This confirms previous findings (Ha, 1988; Gitsaki, 1996; Bonk, 1999). It is, however, evident that with the established trend of correlation levels in the literature, proficiency itself would not be an extremely effective predictor of collocational proficiency, since there is a significant amount of error in the regression. It does seem to be true, as was claimed by Howarth (1996), that individual variation plays a large part in this domain of language knowledge. This fact is made apparent upon examination of the distance from the ideal line of regression of many examinees' test scores (Figure 6). While there do not seem to be learners in this administration who obtained high proficiency and low collocations test scores, or low proficiency and high collocations test scores, the middle area of the grid does illustrate quite a bit more variation. A candidate with a score of 35 on this proficiency test may just as well score near the bottom in collocations proficiency as near the top. However, a score of 45 on the proficiency test virtually guarantees that a candidate's collocations knowledge will be near or above the mean for the whole population. This leads us to speculate that, if there are indeed great individual differences in collocational knowledge which are not predictable by level of proficiency, perhaps these are reflective of underlying differences in the ability or aptitude to perceive, remember, and recall instances of restricted collocation. Ellis (1996) has claimed that individuals' short-term memory capacity may serve as a general constraint on their ability to learn collocations, formulaic speech, lexical phrases, phonology--indeed, on second language learning in general. The test results reported here may be an illustration of this phenomenon, whereby those examinees with aptitude for becoming near-native in a second language have accordingly displayed an at-least average level of collocational

knowledge. Learners who lack this underlying aptitude may never achieve high levels of L2 performance. Finally, from the absence of low-collocation high-proficiency examinees in this administration of the test, it can be deduced that well-developed collocations knowledge may be one of the last stages of second language acquisition, as has been previously suggested in the literature (e.g., Bahns, 1993).

While this study was not set up to answer the question of whether or not collocations ought to be taught in the classroom, or when and how they might be best introduced, the evidence from this test administration seems to indicate that learners acquire collocational knowledge on their own or with informal instruction only. This assertion of course is contingent upon whether or not collocations are actually taught in classrooms or not, which has not been proven in this or any other study.

Length of Residence and Collocations Ability

LOR correlated significantly with collocations scores in the pilot study, but the level of correlation was low enough to not be meaningful ($r = .39$), and a violation of one assumption of this statistic made the result uninterpretable. It may be ultimately a question of quality rather than quantity, as common sense would suggest. In Bonk (1999) I found that some variables measuring the amount of interaction with NSs correlated significantly with collocations test scores, but that virtually all the variance was accounted for by a proficiency variable in factor analysis; I interpret this to mean that interaction with English NSs is only facilitative of the acquisition of collocations when it makes a direct contribution to proficiency; otherwise, it has little or no effect.

Lower-Proficiency Learners and Collocations Knowledge

In the pilot administration the collocations test did not seem to discriminate well among lower-to-intermediate proficiency learners. It was deduced that the test

did not include a sufficient number of items at their level of ability to ensure adequate discrimination. Accordingly, 20 presumably easier items were added to the pilot test and administered as a part of the main study. Eight of these items ultimately proved to have high item facility scores, so there was some effect on the test. Examination of the information function based on IRT analysis (Figure 8) indicates that there is test information for all the lower-proficiency learners (–2 logits) in this study. Therefore it can be concluded that the lack of discrimination in the pilot administration was mainly due to characteristics of the test itself, and that this problem was diminished in the subsequent administration. Lower-level learners do seem to have some limited knowledge of collocational relationships, and it can be tested accurately as long as they have access to the items, which are necessarily written in English. Translation of prompts might be an alternative, but it was not attempted in this study due to the heterogeneous nature of the examinee population.

Validity Evidence from Factor Analysis

The factor analysis results display a clear pattern of the divergence of collocations test scores from those on more general proficiency measures. This provides evidence for the claim in this study that the test of collocations is measuring a construct other than some general aspect of English language proficiency. The convergence of the three collocations subtest loadings on the same factor also provides some evidence for the construct validity of this measure of second language knowledge, since they were intended to examine various aspects of the same type of knowledge.

Limitations of the Study

One area of potential problems has already been pinpointed earlier, that of schema activation and cultural bias in the test. This variable was not controlled for at

BEST COPY AVAILABLE

all in this test administration, and may have had significant impact on candidates' test scores, efforts should be made to reduce this source of variance from prompts in collocations tests. Another potential problem was difference in target varieties: it was assumed that the examinees in this study had mainland American English as their target, but this was not verified; if they had another variety as their target, it would be understandable that their scores would be low on collocations measures.

Factor analysis, generalizability theory, and IRT analysis are very powerful statistics when used with appropriate data sets, but results may be misleading when applied to as small a sample as was reported in this study. Therefore results must be approached with caution, and could hopefully be replicated with larger sample sizes in the future.

True English "proficiency" can only at best be approximated by the type of measure used in this study. There is much important knowledge and competence that this type of test overlooks in its measurement of learners, and therefore the term "proficiency" as I have used it is misleading.

Agenda for Further Research

Now that this test of collocations has been described and pilot-tested, further studies can be undertaken to investigate how other measures of second language acquisition relate to it, and how collocational knowledge fits into existing models of L2 competence. Another area which is worth pursuing is the study of collocations acquisition by L2 learners, through tasks designed to make them more aware of how collocation works and what effect it has on the language, as well as through explicit instruction. It has been asserted in this study that learners may be able to and do acquire this knowledge on their own, but it remains to be seen whether different types of instruction can facilitate this learning.

BEST COPY AVAILABLE

There is of course also much work which needs to be done on the knowledge and performance of native speakers themselves in the area of collocations, since not only is there virtually no empirical evidence, but there is very little discussion on their storage in, relationships within, and retrieval from the mental lexicon. For example, the stance taken in this study has been that collocation means that fast and frequently accessed connections are established between lexical elements, but that these items are not stored together. If their access is faster or slower than the already-studied free combinations and idioms, then there would be some evidence to support or refute this view of collocation.

Conclusion

This project represents the first attempt at a comprehensive description of a norm-referenced test of second language learners' knowledge of collocations, an important yet largely undescribed area of linguistic competence. The results reported here suggest that learners at even low-intermediate levels of general proficiency in English (with TOEFL scores of perhaps only 400 or so) have developed some productive knowledge of target language collocations. It has been shown that this knowledge generally increases with proficiency (though there is a great deal of variation from learner to learner in the relationship between these two variables), and it has been suggested that such knowledge may be acquired naturalistically, since it is probably not a frequent focus of attention in the classroom. Though knowledge of target language collocations is not an extremely efficient predictor of general proficiency in the second language, it has been demonstrated that a certain level of proficiency in this domain can be guaranteed if the level of proficiency is known.

In terms of specific testing issues, both cloze- and selected response-type items have been shown to be relatively reliable ways of measuring this area of

language knowledge, and items of these types were found to be easy to construct, validate, and score. Collocations involving prepositions associated with verbs were not reliably measured in this study, but the other two types functioned adequately as norm-referenced measures of the construct. Factor analysis and the results of a generalizability study concurred in providing evidence that the three subtests investigated here measured the same construct, and that this construct was something not covered by TOEFL-like measures of proficiency.

I have claimed above that collocational knowledge is an important component of one's lexical knowledge in general, and that it has an impact on many aspects of language processing, comprehension, and use. Though they are generally not in use at this time, tests of collocational knowledge could provide language professionals and researchers with potentially valuable information on the lexical relations knowledge of their learners, since collocational knowledge differs from other types of written language proficiency and can be reliably and quickly tested. It can be as conversational or as educated a test as desired, since collocations exist throughout registers and language varieties; indeed, these are defined in part by the existence of specialized collocations. Collocations testing may even provide clues to eventual ultimate attainment in the L2, since it acts as a constraint on "grammatical" language production. However, it is perhaps the practicality of collocations testing that is its strongest point, however. Long prompts are not needed; reliable items can be easily developed and validated by NSs; and there are thousands and thousands of potential items available to testers in every language. The main concern of such tests is that the collocations items exactly match the target varieties of the examinees, since any divergence from this will be strongly reflected in invalid test scores.

BEST COPY AVAILABLE

REFERENCES

Aijmer, K., and Altenberg , B. (Eds.). (1991). English corpus linguistics - Studies in honour of Jan Svartvik. London and New York: Longman.

Aitchison, J. (1987). Reproductive furniture and extinguished professors. In R. Steele & T. Threadgold (Eds.), Language topics: Essays in honour of Michael Halliday Vol. 2 (pp. 24-49). Amsterdam: John Benjamins.

Bahns J., and Eldaw, M. (1993). Should we teach EFL students collocations? System 21, 101-114.

Bahns, J. (1993). Lexical collocations: A contrastive view. ELT Journal 47 (1), 56-63.

Benson, M., Benson, E., and Ilson, R. (1986). The BBI combinatory dictionary of English. Amsterdam and Philadelphia: John Benjamins.

Biskup, D. (1992). L1 influence on learners' renderings of English collocations: A Polish German empirical study. In P. Arnaud & H. Béjoint (Eds.), Vocabulary and applied linguistics (pp. 85-93). London: Macmillan.

Bonk, W. (1999). Japanese university students' knowledge of English collocations. The Journal of Kanda University of International Studies 11, 411-448.

Bonk, W. (1995). Investigating collocational ability among ESL learners. Unpublished manuscript, University of Hawai'i.

Brennan, R. (1983). Elements of generalizability theory. Iowa City, IA: American College Testing Program.

Brown, D. (1974). Advanced vocabulary teaching: The problem of collocation. RELC Journal 5, 1-11.

Ellis, N. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. Studies in Second Language Acquisition 18, 91-126.

BEST COPY AVAILABLE

Frauenfelder, U., and Tyler, L. (1987). The process of spoken word recognition: An introduction. In Cognition 25, 1-20.

Gitsaki, C. (1996). English collocations in second language acquisition. Unpublished Ph. D. dissertation, University of Queensland, Australia.

Ha, M. (1988). Acquisition of collocations: Investigation of relative difficulty among different types of collocations. Unpublished manuscript, University of Hawai'i.

Hambleton, R., Swaminathan, H., and Rogers, H. (1991). Fundamentals of Item Response theory. Newbury Park, California: Sage Publications.

Hatch, E., and Lazaraton, A. (1991). The research manual: Design and statistics for applied linguistics. New York: Newbury House.

Howarth, P. (1998). Phraseology and second language proficiency. Applied Linguistics 19 (1), 24-44.

Howarth, P. (1996). Phraseology in English academic writing. Tübingen: Max Niemayer Verlag.

Hussein, R. (1991). Collocations: The missing link in vocabulary acquisition amongst EFL learners. In Papers and Studies in Contrastive Linguistics Vol. 26, (pp. 123-136). Poznan: Adam Mickiewicz University.

Irujo, S. (1986). Don't put your leg in your mouth: Transfer in the acquisition of idioms in a second language. TESOL Quarterly 20, 287-304.

Kellerman, E. (1986). An eye for an eye: Crosslinguistic constraints on the development of the L2 lexicon. In E. Kellerman & M. Sharwood Smith (Eds.), Crosslinguistic influence in second language acquisition. Oxford: Pergamon Press.

BEST COPY AVA

Kennedy, G. (1990). Collocations: Where grammar and vocabulary teaching meet. In S. Anivan (Ed.), Language teaching methodology for the nineties. (pp. 215-232). Singapore: SEAMEO Regional Language Centre.

Kjellmer, G. (1991). A mint of phrases. In K. Aijmer & B. Altenberg (Eds.), English corpus linguistics - Studies in honour of Jan Svartvik (pp. 111-127). London: Longman.

Mackin R. (1978). Words shall be known by the company they keep. In P. Strevens, (Ed.), In honour of A. S. Hornsby (pp. 149-165). Oxford: Oxford University Press.

Marton, W. (1977). Foreign vocabulary learning as problem no. 1 of language teaching at the advanced level. Interlanguage Studies Bulletin 2, 33-57.

McCarthy, M. (1990). Vocabulary. In the series C. Candlin & H. Widdowson (Eds.), Language teaching: A scheme for teacher education. Oxford: Oxford University Press.

Mislevy, R. and Bock, R. D. (1990). Item analysis and test scoring with binary logistic models. Mooresville, IN: Scientific Software, Inc.

Mislevy, R. and Bock, R. D. (1992). Bilog 3.04. Computer program. Mooresville, IN: Scientific Software, Inc.

Nattinger, J., and DeCarrico, J. (1992). Lexical phrases and language teaching. Oxford: Oxford University Press.

Neter, J., Wasserman, W., and Kutner, M. (1990). Applied linear statistical models. Homewood, IL: Irwin.

Oppenheim, N. (1993). An investigation of recurrent sequences in second language production. Unpublished master's thesis, University of Hawaii.

BEST COPY AVAILABLE

Pawley, A., and Syder, H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. Richards & R. Schmidt (Eds.), Language and communication (pp. 191-225). London: Longman.

Peters, A.M. (1983). The units of language acquisition. Cambridge: Cambridge University Press.

Richards, J. (1976). The role of vocabulary teaching. TESOL Quarterly 10 (1), 77-86.

Riopel, W. (1984). The effect of two methods of ESL instruction on the learning of verb plus particle collocations. Unpublished master's thesis, University of Hawai'i.

Rost, M. (1994). Second language processing: Listening. In R. Asher (Ed.), 1994. Encyclopedia of language and linguistics (pp. 3778-3782). Pergamon Press.

Sinclair, J. (1991). Corpus, concordance, collocation. Oxford: Oxford University Press.

Yorio, C. (1989). Idiomaticity as an indicator of second language proficiency. In K. Hyltenstam & L. Obler (Eds.), Bilingualism across the lifespan (pp. 55-72). Cambridge: Cambridge University Press.

Zhang, X. (1993). English collocations and their effect on the writing of native and non-native college freshmen. Unpublished doctoral dissertation, Indiana University of Pennsylvania.

Zimmerman, C. (1993). Teachers' perceptions of and strategies with lexical anomalies. Paper presented at TESOL conference, April 1993, Atlanta, Georgia.

BEST COPY AVAILABLE

Appendix A

Put the **verb** which best completes the sentence in the blank. You have 8 minutes.

EXAMPLE He was accused of __spreading__ rumors about the other employees.

1. I'm sorry, but I can't give it to you at that price; I'm trying to _____ a business here, not a charity organization.
2. She _____ the subject every time I asked about her boyfriend; she obviously didn't want to talk about him.
3. I _____ lots of pictures on my vacation, but none of them came out nice.
4. Don't eat chocolate before dinner, it'll _____ your appetite.
5. If you don't make back-up copies of all the files on your computer, you'll be _____ the risk of losing all your data during a power failure.
6. I _____ a bad dream last night.
7. Punk rockers dye their hair red and green because they want other people to _____ attention to them.
8. It doesn't _____ any difference if you go left or right here, the two streets will meet again after only 1 mile.

9. The meeting will be _____ on Friday in room 1302.
10. I've been _____ a diary for over ten years now; I love to look back at it now and see what I was thinking about.
11. I like to _____ a nap every afternoon after lunch.
12. My husband would never _____ such a serious decision without first discussing it with me.
13. We all _____ the feeling that she didn't really want to come to our party.
14. She _____ me a big favor and took care of the kids for the afternoon when I had to go to the doctor's office.
15. Whenever he gets drunk, he ends up _____ a spectacle of himself.
16. Please _____ your feet before you come into the house, your shoes look dirty.
17. Orimco, the company that made these chemicals, _____ __ out of business a year ago.

Fill in the blanks with **prepositions** to complete the sentence. The meaning of the verbal expression is in parentheses at the end of the line. You have 8 minutes.

EXAMPLE: Hey, that's a nice leather jacket; mind if I try it __on__? **(to wear)**

18. We ordered wine and two appetizers, and the bill came _____ $78. **(to total)**

19. Don't worry, you can depend _____ me! **(to trust)**

20. I'm sorry I can't drive you all the way to school, but I can drop you _____ at the bus stop if you like. **(to drive someone to a place)**

21. It's taken me more than a month to get _____ this cold. **(to recover)**

22. They set _____ a bomb in the busy marketplace, killing hundreds of people. **(to make explode)**

23. The criminals who held _____ the Bank of Hawaii last week still have not been caught. **(to rob)**

24. Sylvester Stallone's new movie is scheduled to come _____ on April 19. **(to appear)**

25. My doctor told me I had to give _____ alcohol and cigarettes if I want to stay healthy. **(to stop using)**

26. Patty was Ron's girlfriend for a long time, but they broke ____ __ two weeks ago. **(to end a relationship)**

27. Many of the birds in the area were killed _____ by local hunters. **(to exterminate)**

28. I've finished writing up the minutes of the last meeting; would you please look them _____ and see if I've made any mistakes? **(to check)**

29. We've been talking about this for over an hour now; let's move _____ to another topic. **(to go)**

30. My father never studied English formally, he just picked it ___ __ when he came here ten years ago. **(to learn)**

31. Other kids always pick _____ her because she's so overweight. **(to tease)**

32. He takes _____ his father, he has the same hair and eyes. **(to look like)**

33. She told me that she wanted to go to Alaska and work on a fishing boat for the summer, but I talked her _____ of it. **(to convince not to do)**

34. Grandfather has been feeling very sad, let's go to his house and cheer him _____ . **(to make happier)**

BEST COPY AVAILABLE

Each of the four sentences is using the underlined verb in a different way. One of them is not really a correct usage of that word. **Circle the letter corresponding to the least acceptable sentence**. You have 10 minutes.

35. a. You are covered by your doctor's diagnosis.
    b. Another soldier covered Peter with a machine gun as he started forward.
    c. With this insurance policy, we will cover you in case of any accident.
    d. Wow, we've been driving fast! We covered almost three hundred miles in one day.

36. a. After his wife divorced him and his son died last year, Mark took a heart attack.
    b. Our plane took off two hours late due to heavy snow on the runways.
    c. The robbery took place at about 3:30 A.M.
    d. Children have to be taught how to take turns, or they will continue to be selfish.

37. a. Don't put the blame on me!
    b. Can you put your signature at the bottom of this letter, please?
    c. It's OK, I think I can put the risk.
    d. 25 convicted murderers were put to death in Texas last year.

38. a. I can see up your plans, you want to take my house away from me!
    b. Her whole family was at the train station to see her off.
    c. She saw to it that the animals were fed at the same time every day.
    d. Ever since I got hit on the head with that rock, I've been seeing double.

39. a. I didn't leave my apartment until after two o'clock.
    b. My Aunt Janice left me this beautiful tea set.
    c. Roberta left her husband for another man.
    d. Hitler left his mind in the days before his suicide.

40. a. Are the Johnsons throwing another party?
    b. She threw him the advertising concept to see if he liked it.
    c. The team from New Jersey was accused of throwing the game.
    d. The new information from the Singapore office threw the meeting into confusion.

41. a. Let's drop the subject. I don't want to talk about it any more.
    b. He kept dropping hints that he wanted a job, too.
    c. Bob said he would bring the computer over to my house, but I dropped his offer.
    d. She didn't like the professor, so she decided to drop the class.

42. a. I don't know if I can stand another winter in this place.
    b. She stood a good chance of winning the race.
    c. I've already explained where I stand on that issue.
    d. He stood three roses in the vase.

43. a. They play the radio really loud, even late at night.

BEST COPY AVAILABLE

b. Money plays a large part in politics, especially when it comes to rich countries.

c. What movies are playing at the Cinemat tonight?

d. This book played great impact on the intellectuals of Eastern Europe.

44. a. I don't think you can pick a conclusion from this data.

b. Tom is always trying to pick fights with Ronnie.

c. His pockets were picked while standing in line at the theater.

d. Picking berries in the woods is a great way to spend a Sunday afternoon.

45. a. They got big success when they finally discovered how to market their products in Japan.

b. If you get good results from this medication, let me know.

c. He got cancer as a result of his thirty years of smoking cigarettes.

d. We got to New York on the same day that the big snowstorm began.

46. a. He didn't feel that a videocassette would make a very good present, so we bought something else.

b. I made the table while Mom finished cooking dinner.

c. Seven and nine make sixteen.

d. The train crossing signals are starting to flash - do you think we can make it?

46

47. a. I'm not sure if they do repairs in that shop or not.

b. We did our best, but it wasn't good enough to win.

c. What profession does he do?

d. She didn't do very well on the test.

48. a. We're going to run some tests on your mother to see if the accident affected her brain.

b. All the newspapers in the country ran the story on their front pages.

c. A car ran over my cat yesterday.

d. You need to run more vitamins and minerals in your diet.

49. a. Could you please move your car - it's blocking my driveway.

b. I was very moved by her latest film.

c. The situation seems very bad now for you, but it may move at any time.

d. The neighborhood was getting too dangerous, so they moved.

50. a. Bob pulled a terrible trick with my computer, and now I can't get it to work again.

b. A little kid pulled a gun on me right outside my apartment last night!

c. I want to make sure that everybody in this company is pulling their own weight.

d. She's never pulled anything like this before.

47

Table 1

Collocations Test Scores (N = 98)

	Collocations Total	Verb-Object	Verb-Prep	Figurative Verbs
k	50	17	17	16
M	25.28	8.66	8.78	7.84
Median	26	9	9	8
Mode	27	11	10	10
SD	7.28	3.12	2.34	3.0
SEM	3.03	1.75	1.70	1.90
K-R 20	83	.69	.47	.61
Mean IF	.51	.51	.48	.53
Skewness	-.09	-.20	-.24	-.15
Kurtosis	-.59	-.46	-.49	-.04
Range	32	14	11	14

Table 2

Proficiency Test Scores (N = 98)

	Proficiency Total	Grammar	Vocabulary	Reading
k	49	20	15	14
M	37.26	15.78	12.17	9.31
SD	7.15	2.98	2.20	3.04
SEM	2.74	1.72	1.42	1.68
K-R 20	.85	.67	.58	.69
Skewness	-.41	-.755	-.627	-.375
Kurtosis	-.59	-.524	.136	-.711
Range	30	14	10	12

Table 3

Collocations Test Item Analysis

Verb-object				Verb-preposition				Figurative verbs			
Item	IF	ID	pbi	Item	IF	ID	pbi	Item	IF	ID	pbi
1	.17	.15	.18	18	.20	.11	.14	35	.45	.47	.34
2	.56	.50	.43	19	.98	.06	.27	36	.79	.46	.48
3	.77	.27	.27	20	.70	.14	.17	37	.71	.56	.54
4	.14	.36	.46	21	.49	.50	.46	38	.27	.33	.33
5	.36	.48	.45	22	.10	.21	.31	39	.63	.52	.43
6	.79	.14	.13	23	.18	.18	.28	40	.34	.23	.16
7	.74	.33	.31	24	.35	.38	.35	41	.45	.22	.27
8	.66	.77	.68	25	.96	.06	.17	42	.65	.33	.36
9	.69	.39	.39	26	.70	.39	.30	43	.54	.25	.24
10	.22	.39	.45	27	.21	.01	.06	44	.29	.11	.12
11	.77	.45	.45	28	.31	.14	.24	45	.39	.60	.48
12	.78	.42	.45	29	.62	.43	.42	46	.38	.04	.11
13	.85	.23	.32	30	.83	.14	.16	47	.43	.35	.35
14	.52	.50	.41	31	.35	.33	.35	48	.66	.59	.49
15	.26	.36	.37	32	.42	.32	.26	49	.59	.62	.50
16	.16	.21	.22	33	.47	.38	.35	50	.28	.26	.24
17	.24	.42	.43	34	.90	.23	.35				
M	.51	.43	.38		.48	.25	.27		.53	.37	.34

Note. ID and point-biserial coefficients were calculated based on collocations test score totals.

Table 4

Inter-Test Correlation Matrix (N = 98)

	Verb-object	Verb-prep	Fig. Verbs	Grammar	Vocabulary	Reading	Colls total	Proficiency total
Verb-object	1							
Verb-prep	.64	1						
Fig. Verbs	.63	.55	1					
Grammar	.46	.37	.31	1				
Vocabulary	.48	.41	.38	.64	1			
Reading	.54	.51	.56	.62	.65	1		
Colls total	.89	.82	.86	.45	.50	.63	1	
Proficiency total	.57	.50	.49	.88	.85	.88	.61	1

Note. All correlations are significant at $p < .05$, one-tailed test.

Table 5

IRT-Based Item Thresholds, Associated Error Values, Slopes, and Item Reliability Estimates

Item	Threshold	Error	Slope	Maximum Information	Reliability
1	1.862	.543	.555	.2229	.1049
2	-0.177	.173	.776	.4352	.2445
3	-1.276	.374	.589	.2508	.1399
4	1.17	.205	1.403	1.4226	.3278
5	0.525	.168	.946	.6462	.2918
6	-1.86	.626	.427	.1315	.0784
7	-1.018	.317	.644	.2998	.1683
8	-0.345	.100	2.003	2.8992	.5350
9	-0.738	.246	.725	.3799	.2083
10	0.98	.219	1.035	.7733	.2800
11	-0.919	.225	.912	.6003	.2540
12	-0.88	.199	1.055	.8049	.2956
13	-1.45	.357	.807	.4707	.1814
14	-0.023	.167	.800	.4626	.2547
15	0.923	.232	.911	.5990	.2534
16	1.874	.537	.582	.2444	.1091
17	0.977	.218	.965	.6721	.2630
18	1.868	.575	.472	.1612	.0881
19	-2.14	.598	1.427	1.4723	.1471
20	-1.148	.415	.461	.1534	.1050
21	0.087	.165	.836	.5052	.2677
22	1.584	.299	1.078	.8398	.2113
23	1.505	.402	.693	.3465	.1529

24	0.681	.224	.687	.3412	.1976
25	-2.426	.720	.887	.5686	.0965
26	-0.977	.322	.554	.2218	.1405
27	2.046	.671	.404	.1177	.0692
28	0.951	.299	.601	.2613	.1573
29	-0.431	.201	.729	.3843	.2215
30	-1.845	.574	.531	.2037	.1009
31	0.736	.247	.615	.2735	.1698
32	0.448	.242	.537	.2086	.1469
33	0.173	.188	.683	.3375	.2088
34	-1.693	.394	.903	.5886	.1733
35	0.289	.232	.558	.2253	.1575
36	-0.954	.218	1.000	.7229	.2743
37	-0.595	.160	1.202	1.0431	.3591
38	1.031	.288	.714	.3682	.1892
39	-0.433	.181	.818	.4836	.2535
40	1.034	.376	.432	.1348	.0978
41	0.286	.224	.567	.2326	.1611
42	-0.561	.215	.725	.3795	.2155
43	-0.176	.264	.469	.1586	.1217
44	1.357	.462	.440	.1400	.0939
45	0.437	.173	.879	.5587	.2744
46	0.869	.395	.377	.1026	.0811
47	0.336	.195	.691	.3452	.2094
48	-0.491	.166	.963	.6701	.2991
49	-0.261	.163	.890	.5724	.2839
50	1.204	.368	.543	.2127	.1290

Table 6

Item-Level Fit Statistics for the IRT Model

Item	Chi-square	Degrees of freqeuncy	Probability
1	1.3	1	.2495
2	2.5	3	.4789
3	7.4	2	.0243
4	7	0	1.000
5	1.9	2	.3884
6	4.9	2	.0839
7	2.4	3	.4898
8	0	0	1.000
9	1.1	3	.7801
10	1.0	1	.3078
11	3.4	2	.1793
12	2.8	2	.2500
13	1.3	1	.2492
14	3.8	3	.2801
15	1.5	2	.4686
16	.5	1	.4668
17	2.1	1	.1414
18	1.9	2	.3877
19	1.6	0	1.000
20	5.3	4	.2547
21	1.8	3	.6238
22	1.5	1	.2132
23	.9	2	.6431
24	1.7	2	.4355
25	.7	0	1.000

26	6.4	4	.1685
27	7.2	2	.0268
28	6.3	2	.0421
29	4.9	4	.3014
30	2.2	2	.3327
31	4.0	2	.1333
32	9	2	.6363
33	2.1	3	.5557
34	1.1	1	.2864
35	3.3	3	.3455
36	2.0	1	.1529
37	1.4	1	.2389
38	2.0	2	.3727
39	1.9	2	.3970
40	2.2	2	.3274
41	3.8	3	.2789
42	5.2	4	.2705
43	7.2	4	.1228
44	5.9	3	.1144
45	1.9	2	.3850
46	7.9	3	.0483
47	4.5	2	.1045
48	.8	2	.6678
49	1.3	2	.5297
50	6.8	3	.0779

Table 7

Collocations Test Generalizability Study Variance Components

Variance Contributors	Variance Component Estimates
Person (p)	.0183
Subtest (s)	-.0055[a]
Items:Subtest (i:s)	.0011
Person x Subtest (p x s)	.0004
Person x Item:Subtest (p x (i:s))	.0037
δ	.004
Δ	.005

Note. Estimates of variance components are based on scores of 0 (incorrect) or 1 (correct).

[a]This negative value was rounded to 0 for analysis, after Brennan (1983).

Table 8

Dependability Study of Collocations Test

Total subtests	Items per subtest	Total items	G - coefficient	δ	Δ
3	5	15	.60	.012	.016
1	16	16	.60	.012	.016
5	5	25	.71	.007	.010
3	10	30	.74	.006	.008
2	16	32	.75	.006	.008
2	24	48	.81	.004	.005
*3	*16	*48	*.82	*.004	*.005
6	8	48	.82	.004	.005
5	10	50	.83	.004	.005
3	20	60	.84	.003	.004
4	16	64	.85	.003	.004
3	25	75	.87	.003	.004
5	15	75	.87	.003	.003
4	25	100	.90	.002	.003
5	20	100	.90	.002	.003

Note. Asterisks indicate data corresponding to the actual administration reported in this study.

Table 9

Factor Loadings and Communality Values for Collocation and Proficiency Subtests: Oblique Solution Primary Pattern Matrix

Subtest	Factor 1	Factor 2	h^2
Verb-object collocations	.781	.144	.562
Verb-preposition collocations	.835	.014	.463
Figurative verb collocations	.891	-.054	.496
Grammar proficiency	-.086	.938	.500
Vocabulary proficiency	.019	.874	.523
Reading proficiency	.325	.649	.601

Table 10

Proportionate Variance Contributions of the Two Factors - Oblique solution

	Direct	Joint	Total
Factor 1	.409	.203	.611
Factor 2	.385	.003	.389

Table 11

Analysis of 24 Student Error Responses to verb-Preposition Subtest Items

Item	Target Response	Mean IF (N=98)	Most popular error responses and number of tokens	Other error responses and number of tokens
18	come to	.20	up 7, out 5	of 1, at 2, up to 1, with 1, in 1
19	depend on	.98	-	-
20	drop off	.70	down 2, out 2	by 1, on 1
21	get over	.49	off 3	through 1, out 1, up 1, into 1, on 1
22	set off	.10	up 17	on 3, down 2, for 1
23	hold up	.18	on 3, out 3	into 2, at 1, in 2, on 1, off 2, from 1, of 1
24	come out	.35	up 13	in 2, on 1, across 1
25	give up	.96	off 1	
26	break up	.70	apart 2, off 2, down 2	out 1, away 1
27	kill off	.21	up 7	over 3, up 3, down 2, after 1
28	look over	.31	up 7, through 3	for 1, into 1, out 1
29	move on	.62	out 5	up 4
30	pick up	.83	on 2	out 1
31	pick on	.35	out 4, up 3	over 1, at 1, down 2
32	take after	.42	on 2, as 2, of 2	for 1, in 1, from 1, like 1
33	talk out	.47	off 2, over 2	against 1, about 1, with 1
34	cheer up	.90	-	-

Note. Dashes (-) indicate that all student responses were correct for a given item. Some blanks were left empty by students; these are not included in the table.

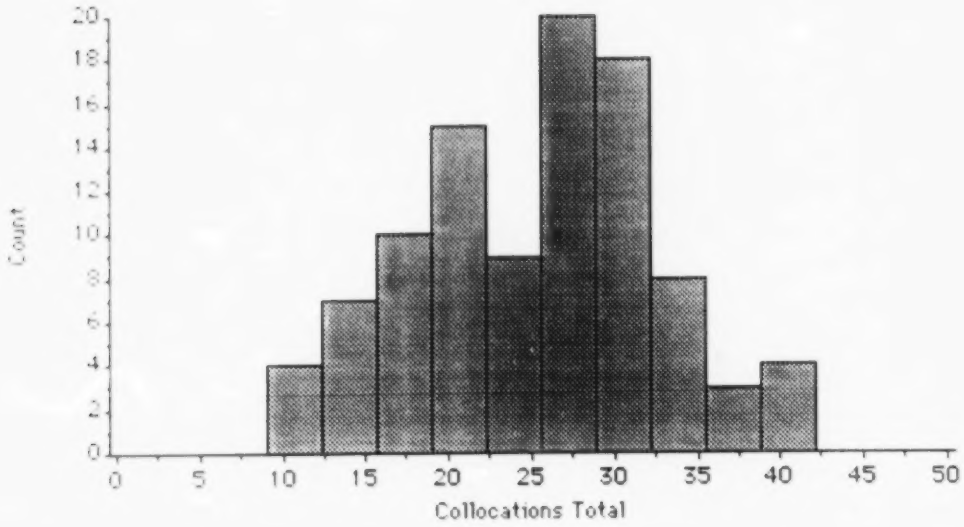


Figure 1. Frequency distribution of collocations test score totals.

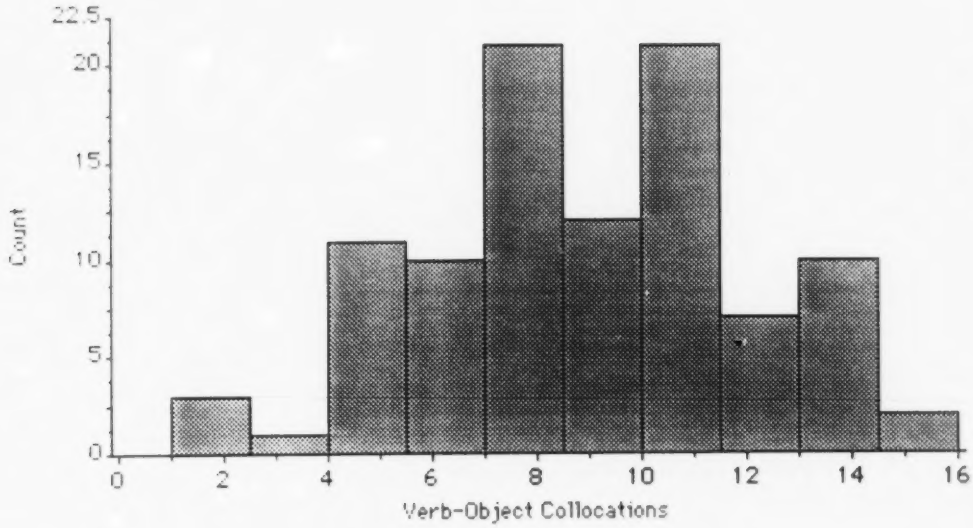


Figure 2. Frequency distribution of collocations verb-object subtest score totals.

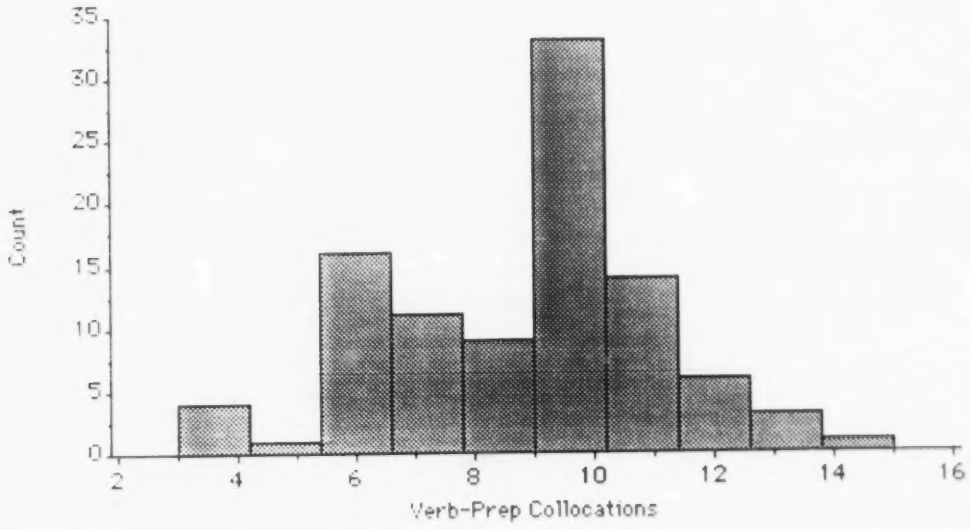


Figure 3. Frequency distribution of collocations verb-preposition subtest score totals.

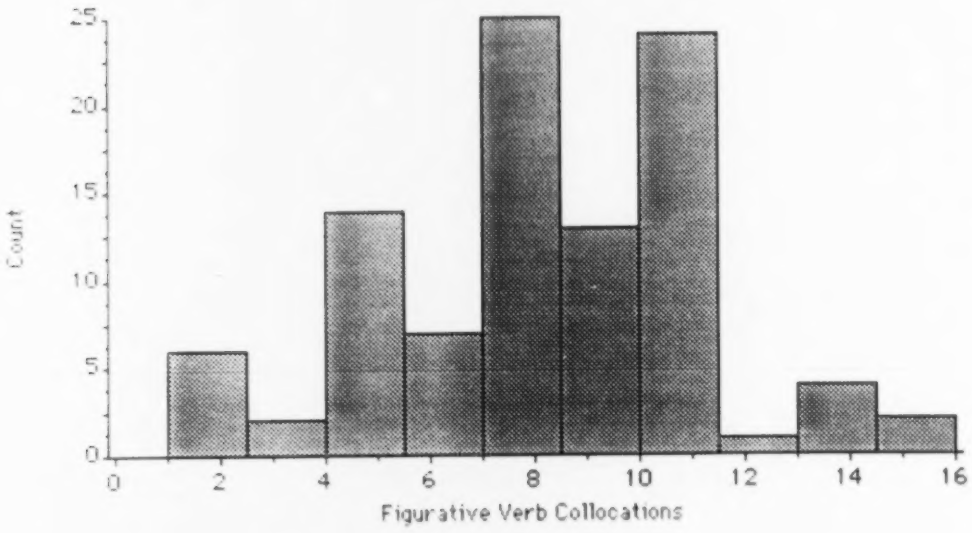


Figure 4 Frequency distribution of collocations figurative verbs subtest score totals.

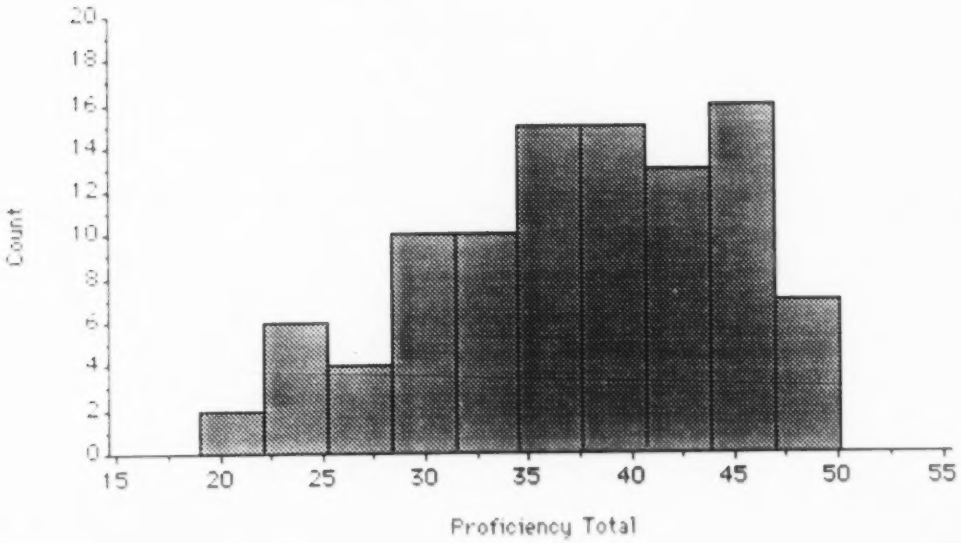


Figure 5. Frequency distribution of proficiency test score totals.

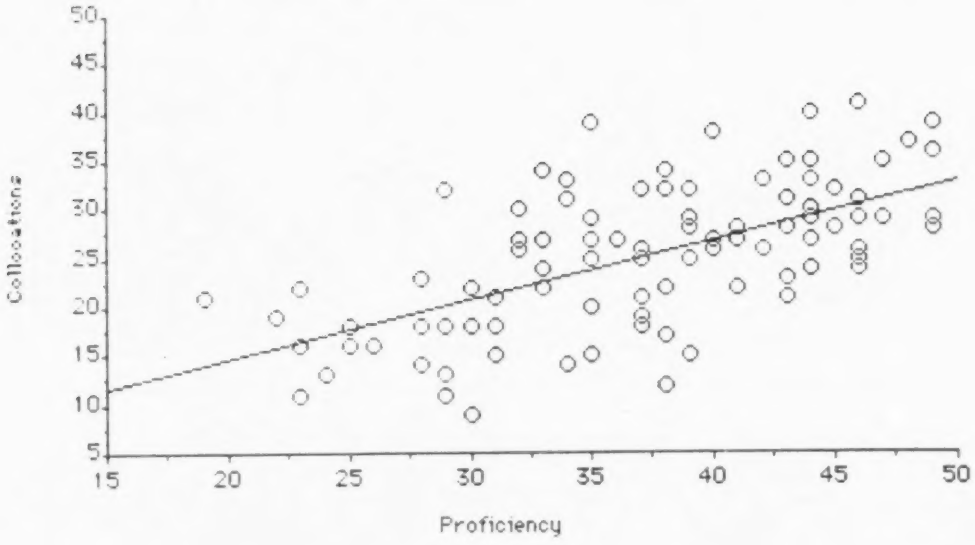
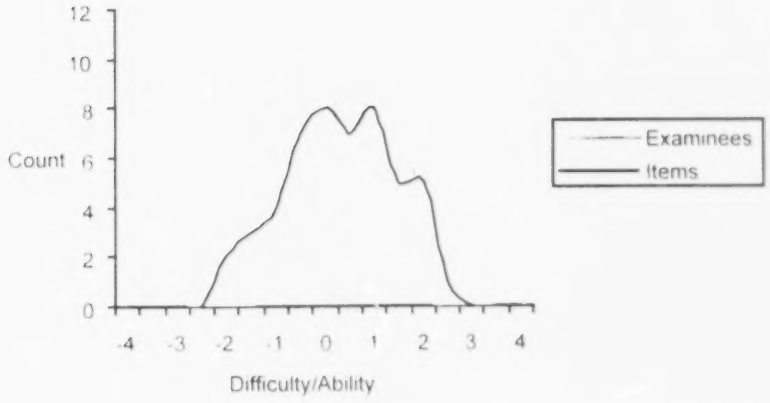


Figure 6. Scatterplot of simple regression collocations - proficiency test scores with ideal line of regression.



Note. Each increment of one represents 3 examinees; items are as shown on scale.

Figure 7. Map depicting match of IRT item difficulty to candidate ability on collocations test.

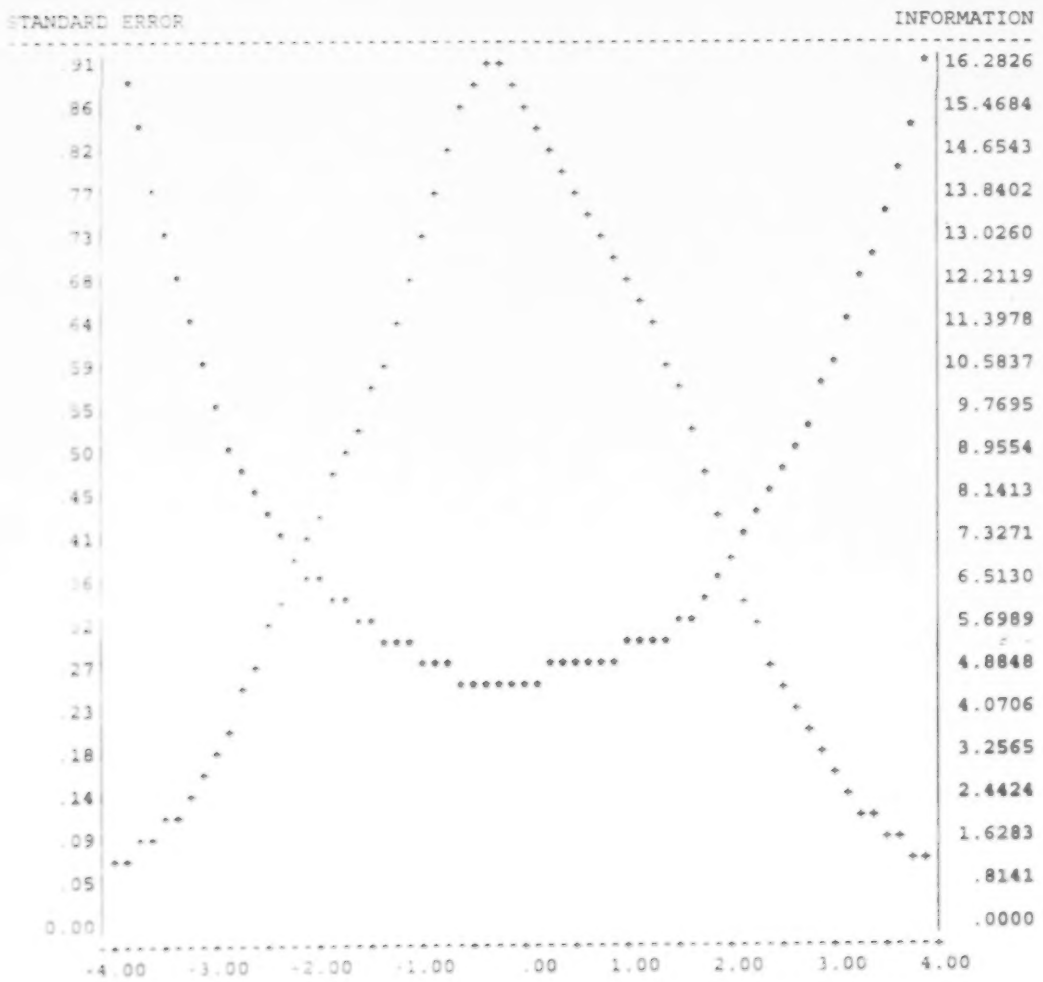


Figure 8. Information and standard error amounts for collocations test across ability levels.

BEST COPY AVAILABLE

Footnotes

[1] I am indebted to Prof. Kate Wolfe-Quintero for this observation.

[2] This test administration is fully described in Bonk (1995).

[3] Two items, #2 and #3 on the verb-object subtest, had a 5-5 split in correct answers among NSs, so it was decided to keep mark both answers as correct responses in NNS testing.

[4] This version of the collocations test was later administered to a group of 193 Japanese university students with much less international study and travel experience; test performance and correlation levels with proficiency scores were similar, suggesting that this test is usable in a FL context as well. The test administration is described in detail in Bonk (1999).

[5] In Bonk (1999) I reported a similar level of correlation for a Japanese L1-homogeneous group of 193 examinees using the same tests: $r = .67$ using raw scores, and $r_{ca} = .82$ using IRT-derived scores on both measures.

[6] In some cases it was difficult to decide how to code a response, such as when the examinee used a delexicalized verb such as take or have. Such responses were coded as appropriate in this study.

BEST COPY AVAILABLE